Empirical Methods in Natural Language Processing

Peking University, 2023

Homework 2: Due on Sunday, April 16 at 11:59 p.m.

Instructions

Please read these instructions to ensure you receive full credits on your homework.

- Submit your homework as a **zip** file through **Course**, which should include one report in PDF, your source code and prediction results on the test set.
- Any coding language is acceptable, but your code should be **your own**. Do NOT submit Jupyter or other notebooks, but the original source code only. We provide a code sample for evaluation in Python.
- You should write your report in **English** and submit it in PDF. We recommend using the official ACL style template for your report (https://github.com/acl-org/acl-style-files).
- Your code should be paired with a README file describing dependencies, code structures, etc.
- There is no need to submit the data you used and the model weights. Your grade will be based on the contents of the report and the source code.

Late submission policy

- Late homework will have 5% deducted from the final grade for each day late.
- **NO** submission will be accepted after April 23, a week after the due date. It is non-negotiable.

- Your submission time will be based on the time of your last submission to Course. Therefore, do NOT resubmit after midnight on the due date unless you are confident that the new submission is significantly better to overcompensate for the points lost.
- You can resubmit as much as you like, but each time you resubmit please be sure to upload all files you want graded!
- The number of points deducted will be rounded to the nearest integer.

Problem Description

In this homework, you will implement models for an online sexism detection task EDOS (Explainable Detection of Online Sexism, SemEval-2023 Task 10).

Sexism is a growing problem online. It can inflict harm on women who are targeted, make online spaces inaccessible and unwelcoming, and perpetuate social asymmetries and injustices. Automated tools are now widely deployed to find, and assess sexist content at scale but most only give classifications for generic, high-level categories, with no further explanation. ***Flagging what is sexist content and also explaining why it is sexist*** *improves interpretability, trust and understanding of the decisions that automated tools use, empowering both users and moderators.*

There are 14,000 instances in the training set, 2,000 in the development set, and 4,000 in the test set.

The original task contains three hierarchical subtasks:

SUBTASK A - Binary Sexism Detection: a two-class (or binary) classification where systems have to predict whether a post is sexist or not sexist.

SUBTASK B - Category of Sexism: for posts which are sexist, a four-class classification where systems have to predict one of four categories: (1) threats, (2) derogation, (3) animosity, (4) prejudiced discussions.

SUBTASK C - Fine-grained Vector of Sexism: for posts which are sexist, an 11-class classification where systems have to predict one of 11 fine-grained vectors.

You will implement models on **subtasks A&B** in this homework.

There is **no constraint** on the methods and data you use. You can implement your own model from scratch, finetune on pretrained language models, or use existing toolkits. You may use the labels of subtask C in the training set to assist with training. The unlabeled data provided by the task constructors, including 1M Reddit entries and 1M Gab entries, may also be a great resource. **Please clearly describe the methods and data you use in the report.**

We will not simply grade your homework based on the model performance, but consider the resources you use and the novelty of your method.

Useful links

Task description paper: https://arxiv.org/abs/2303.04222

Github (including the unlabeled data): https://github.com/rewire-online/edos

Contact

If you have any question about this homework, please email TA via lxlisa@pku.edu.cn .